Mrs. Swati Amit Motugade

STATISTICS WORKSHEET 6

MCQ

1. d

2. a

3. a

4. c

5. c

6. a

7. c

8. b

9. b

10. What is the difference between a boxplot and histogram?

Ans.

1. Box Plot:

A box plot is a form of data visualization that is used to display several aspects of a data set. A segmented box represents the median, upper quartile (Q3), and lower quartile (Q1). A quartile is a segment of data that composes 25% of observed responses, while the median is the data value that represents the middle data value. Lower and higher limits are displayed at the end of whiskers, which are lone segments that extend from the main box. Boxplots may also depict values that are far outside of the normal range of responses (outliers).

2. Histogram:

A histogram is a graphical representation of the spread of data points. Values are broken up into ranges on the x-axis, and the number of responses is denoted on the y-axis. The total number of responses that occur within each

range is depicted as a bar graph. Histograms are helpful in determining how data is distributed.

A histogram depicts the distribution of a data set. Alternatively, a box plot shows quartile values and the median, as well as clearly depicted outliers.

The only difference between a histogram and a bar chart is that a histogram displays frequencies for a group of data, rather than an individual data point; therefore, no spaces are present between the bars. Typically, a histogram groups data into small chunks (four to eight values per bar on the horizontal axis), unless the range of data is so great that it easier to identify general distribution trends with larger groupings.

Q. 11. How to select metrics?

Ans. 'Good Metrics' can be broadly defined as metrics that show if you're achieving your objectives (the ones you prioritized before). Fundamentally, good metrics have three characteristics.

1. **Good metrics are important** to your company growth and objectives. Your key metrics should always be closely tied to your primary objective. A good metric example might be month-on-month revenue growth or LTV:CAC ratio. 'Important' is somewhat subjective since growth for one company may be centred around revenue while another company may focus more on user growth. The key point is to choose metrics that clearly indicate where you are now in relation to your goals.

2. **Good metrics can be improved.** Good metrics measure progress, which means there needs to be room for improvement. For example, reducing churn by 0.8% or increasing your activation rate by 3%. One exception to this might be customer satisfaction - if you're already at 100%, your team will be focused on maintaining that level instead of improving it.

3 **Good metrics inspire action.** When your metrics are important and can be improved, you and your team will immediately know what to do or what questions to ask. For example, why has our conversion rate dropped? Did we make site changes or test a new acquisition channel? Why is churn increasing? By asking questions you can determine possible causes and work to resolve them right away.

Q. 12. How do you assess the statistical significance of an insight?

Ans. Statistical significance can be accessed using hypothesis testing:
- Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)
- Then, we choose a suitable statistical test and statistics used to reject the null hypothesis
- Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)
- We calculate the observed test statistics from the data and check whether it lies in the critical region

Common tests:
- One sample Z test
- Two-sample Z test
- One sample t-test
- paired t-test
- Two sample pooled equal variances t-test
- Two sample unpoled unequal variances t-test and unequal sample sizes (Welch's t-test)
- Chi-squared test for variances
- Chi-squared test for goodness of fit
- ANOVA (for instance: are the two regression models equals? F-test)
- Regression F-test (i.e., is at least one of the predictors useful in predicting the response?)

Q. 13. Give examples of data that doesn’t have a Gaussian distribution, nor log-normal.

Ans. Any type of categorical data won't have a gaussian distribution or lognormal distribution. Exponential distributions - e.g. the amount of time that a car battery lasts or the amount of time until an earthquake occurs, distributions of income; distributions of house prices; distributions of bets placed on a sporting event.

Q. 14. Give an example where the median is a better measure than the mean.

Ans. For distributions that have outliers or are skewed, the median is often the preferred measure of central tendency because the median is more resistant to outliers than the mean.

Example: If the score of students in a class are 1,2,3,4,20

So, if we calculate the mean of these values, it will be 6.

And the median is 3.

So, median is better or appropriate measure because 20 is much greater than other numbers and because of 20 the mean has come out to 6.

Therefore, sometimes it's better to take median than mean.

Q. 15. What is the Likelihood?

Ans. The likelihood is the probability that a particular outcome is observed when the true value of the parameter is equivalent to the probability mass on it is not a probability density over the parameter. The likelihood should not be confused with which is the posterior probability of given the data.